ICLR 2026

From Narrow to Panoramic Vision

Attention-Guided Cold-Start Reshapes Multimodal Reasoning

Ruilin Luo, Chufan Shi, Yizhen Zhang, Cheng Yang, Songtao Jiang,
Tongkun Guan, Ruizhe Chen, Peng Wang, Mingkun Yang, Yujiu Yang, Junyang Lin, Zhibo Yang

Tsinghua University | Qwen Team (Alibaba Group) | University of Southern California

Table of Contents

Background & Motivation

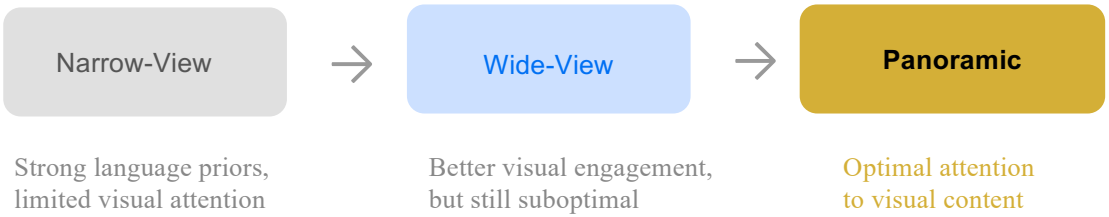
PROBLEM

Multimodal Large Language Models (MLLMs) often fail to fully leverage visual information in reasoning tasks. Cold-start pretraining shows inconsistent improvements.

RESEARCH GAP

Understanding how attention works in multimodal reasoning can guide better training strategies. We discovered attention allocation is critical.

MODEL CATEGORIES



THE COLD-START QUESTION

Does initializing with image-text pairs actually improve visual reasoning?

Key Discovery 1: Visual Attention Score (VAS)

VAS DEFINITION

$$VAS = \Sigma\, Attn(visual\ tokens) \,/\, \Sigma\, Attn(system\ tokens)$$

Measures attention allocated to visual tokens relative to system tokens

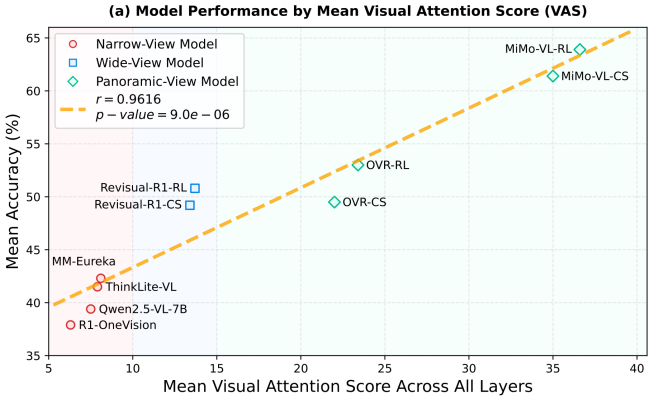
CORRELATION WITH PERFORMANCE

Pearson r = 0.9616

Very strong positive correlation!

TRAINING INSIGHT

- VAS serves as an effective proxy for measuring multimodal reasoning capability
- Models with higher VAS consistently outperform on math, science, and reasoning benchmarks
- This critical insight guided the entire research direction



Key Discovery 2: Lazy Attention Localization

Counter-intuitive finding: Multimodal cold-start FAILS to raise VAS

✗ Multimodal Cold-Start

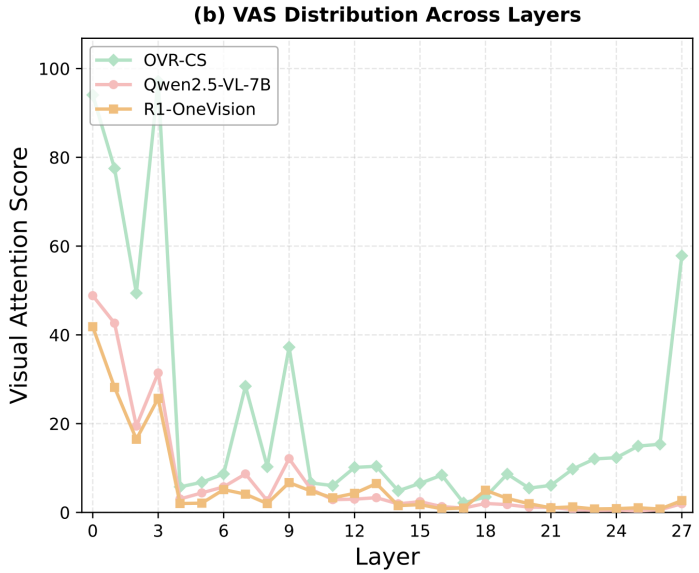
FAILS to raise VAS

Models don't attend to visual tokens even after visual pretraining

✓ Text-only Cold-Start

Clear VAS increase

Model learns to attend to text tokens effectively



ROOT CAUSE ANALYSIS

- Strong Language Priors
- System Token Dominance
- "Seen but Not Attended"

TRAINING IMPLICATION

Simply adding visual data in cold-start is not sufficient - this insight led to AVAR framework

Training-free Interventions for Attention Modulation

Inference-Time Boost for Visual Token Attention in Multimodal Reasoning

PROBLEM

Even with VAS as a metric and understanding of Lazy Attention, how can we improve attention without model retraining?

METHOD

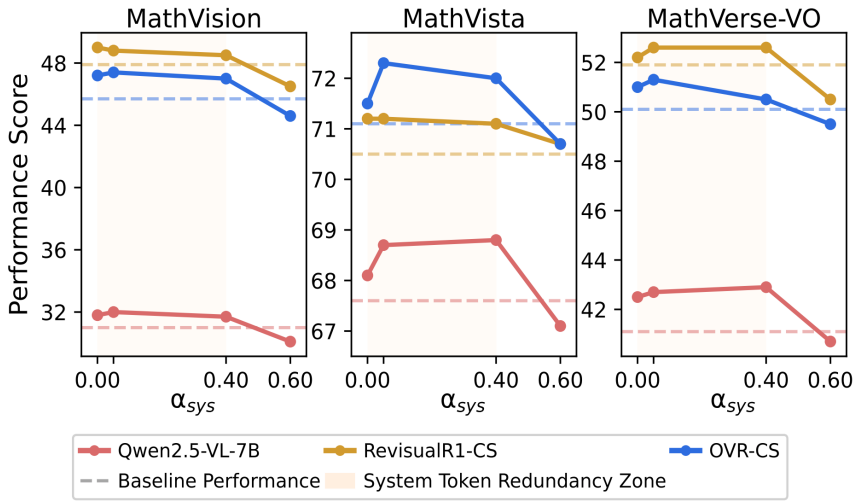
$$Attention\_modified = softmax(Q \cdot K^T + \beta * VisualMask)$$

β controls the strength of visual token boost

SOLUTION

Attention Modulation at Inference Time

Manipulate attention weights during inference

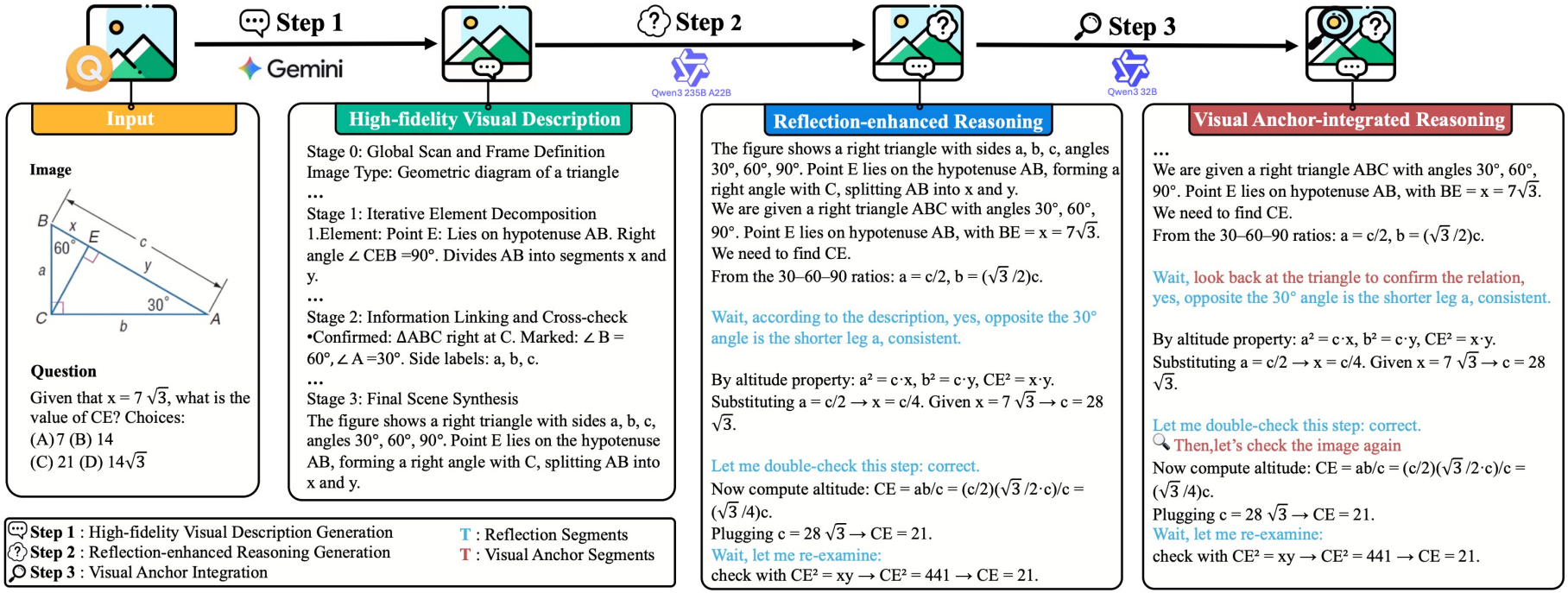


RESULTS

+1-2%

Consistent gains across multiple benchmarks

AVAR Framework: Attention-Guided Visual Anchoring and Reflection



KEY RESULT

7.0% average gain across 7 multimodal reasoning benchmarks on Qwen2.5-VL-7B

MathVista | MathVision | MathVerse-VO | MMMU | MMMU-Pro | MMStar | HallusionBench

Results

	Math Reasoning			Multidisciplinary		Perception		
Model	**MathVista**	**MathVision**	**MathVerse-VO**	**MMMU-VAL**	**MMMU-Pro**	**MMStar**	**Hallusion.**	**Avg.**
Closed-Source								
GPT-4o	63.8	31.2	-	70.7	**54.5**	65.1	56.2	-
Claude-3.7-Sonnet	**74.5**	**58.6**	-	**75.2**	50.1	**68.8**	**58.3**	-
Open-Source General Models								
Qwen2.5-VL-7B	68.2	25.2	41.1	58.1	38.3	62.1	50.7	49.1
InternVL2.5-8B	64.4	22.0	39.5	56.0	38.2	63.2	51.1	47.8
LLaVA-OneVision-7B	58.6	18.3	19.3	48.8	35.5	61.7	47.5	41.4
Llama-3.2-11B-Vision-Instruct	48.6	19.7	18.4	50.7	33.0	49.8	40.3	37.2
Multimodal Reasoning Models								
Mulberry-7B†	63.1	-	42.9	55.0	34.8	61.3	54.1	-
R1-OneVision	64.1	29.9	40.0	49.1	32.2	52.2	46.0	44.8
OpenVLThinker	72.3	25.9	44.6	53.0	42.9	59.5	53.0	50.2
ThinkLite-VL	**75.1**	32.9	45.8	55.5	40.0	65.0	52.3	53.1
MM-Eureka-7B	73.0	26.9	48.1	52.0	42.4	**65.2**	50.7	51.2
Vision-R1†	73.5	-	47.7	56.3	39.6	64.8	51.9	-
VLAA-Thinker-7B	68.0	26.4	48.2	55.7	40.9	64.2	50.9	50.6
Vision-SR1	68.1	26.7	47.1	61.3	**43.8**	64.1	54.3	52.2
Our model								
AVAR-Thinker	74.7	**37.4**	**50.4**	**63.8**	42.9	64.1	**59.5**	**56.1**
Δ over Qwen2.5-VL-7B	+6.5	+12.2	+9.3	+5.7	+4.6	+2.0	+8.8	+7.0

Configuration	**Method Components**			**Benchmark Performance**							
	VARD	**AGTO**	**VARS**	**MathVista**	**MathVision**	**MathVerse-VO**	**MMStar**	**MMMU-VAL**	**MMMU-Pro**	**Hallusion.**	**Avg.**
Baseline (Qwen2.5-VL-7B)				68.2	25.2	41.1	62.1	58.1	38.3	50.7	49.1
	✓			70.6	32.9	43.5	61.1	55.2	38.7	55.3	51.0
	✓	✓		72.0	34.1	44.0	62.8	58.3	39.8	57.2	52.6
AVAR-Thinker	✓	✓	✓	**74.7**	**37.4**	**50.4**	**64.1**	**63.8**	**42.9**	**59.5**	**56.1**

Method	**Benchmark Performance**							**Avg.**
	MathVista	**MathVision**	**MathVerse-VO**	**MMStar**	**MMMU-val**	**MMMU-Pro**	**Hallusion.**	
Baseline (Qwen2.5-VL-7B)	68.2	25.2	41.1	**62.1**	**58.1**	38.3	50.7	49.1
+ R1-OneVision	63.3	26.3	39.7	54.9	49.9	34.6	43.8	44.6
+ OpenVLThinker	68.9	25.3	37.8	58.7	55.7	36.0	54.1	48.1
+ Vision-SR1	67.6	27.9	42.3	46.9	50.7	36.3	42.1	44.8
+ VARD (Ours)	**70.6**	**32.9**	**43.5**	61.1	55.2	**38.7**	**55.3**	**51.0**

Two-Phase Training Pipeline

Conclusion

VAS

Visual Attention Score
r = 0.9616 correlation

Lazy Attention

Root cause of
cold-start failure

Training-free

1-2% gains
without retraining

AVAR

7.0% gain
7 benchmarks

IMPACT

This work provides a fundamental understanding of multimodal attention dynamics and delivers practical solutions that bridge visual pretraining and genuine multimodal reasoning.

Thank You!


清華大學
Tsinghua University
Qwen
USC University of Southern California